Analiza poginulih u prometu

e.math.hr/math_e_article/br17/babic_et_al

Barbara Babić Katarina Bošnjak Nika Kenda Ana Kolić Ivana Kranjec

Sažetak

Koliko puta ste čuli da su žene lošiji vozači od muškaraca ili da su mladi zbog svoje neopreznosti i neiskustva glavni krivci za prometne nesreće? Ovim člankom odlučile smo istražiti govore li i statistički podaci u prilog tim tvrdnjama. Također provjeravamo kakav je utjecaj promjene Zakona o sigurnosti prometa na cestama na broj poginulih, te utječe li obrazovanje vozača na učestalost njihova stradavanja u prometnim nesrećama.
Od 2004. do 2008. godine u Hrvatskoj se dogodilo 307 470 prometnih nesreća u kojima je poginulo 3102 ljudi. Ove zastrašujuće brojke dovoljan su razlog da se ovo istraživanje ne shvati olako.

1Uvod

U posljednjih nekoliko godina u Hrvatskoj je sve izraženiji problem nesigurnosti na cestama i velikog broja prometnih nesreća. Svakodnevno smo okruženi lošim vijestima s prometnica te pokušajima da se promjenama zakona i akcijama MUP-a takvo stanje promijeni. Ponukani time, odlučile smo detaljnije istražiti neke od aspekata te crne statistike.

Točnije, ciljevi ovog rada su:

- ispitati ovisnost smrtnosti po dobnim skupinama o spolu, dobu dana i danu u tjednu
- ispitati ovisnost smrtnosti o stupnju obrazovanja u svim dobnim skupinama
- odrediti očekivanu dob vozača u trenutku nesreće
- provjeriti utjecaj promjene Zakona o sigurnosti prometa na cestama na smrtnost u dobnoj skupini 20 - 29

Prije analize podataka, važno je upoznati se s temeljnim pojmovima korištenima u članku, pa slijedi kratak prikaz glavnih definicija koje se spominju u nastavku.

Statistika je skup ideja i metoda koje se upotrebljavaju za prikupljanje i interpretaciju podataka u nekom području istraživanja te za izvođenje zaključaka u situacijama gdje su prisutne nesigurnosti i varijacije.

Statistička populacija je potpun skup mogućih mjerenja ili podataka o nekom svojstvu koji odgovaraju cijeloj familiji jedinki koju se promatra. U našem slučaju populaciju čine vozači/vozačice koji su poginuli u prometnim nesrećama u razdoblju od srpnja 2004. do lipnja 2009. Podaci su dobiveni iz Državnog zavoda za statistiku, a među ostalim sadržavaju informacije o dobnoj, spolnoj i obrazovnoj strukturi poginulih te o mjesecima, odnosno danima kad su se nesreće dogodile.

Svrha procesa prikupljanja podataka je izvođenje zaključaka o populaciji. Budući da nije uvijek moguće prikupiti sve podatke o području istraživanja, zaključci izvedeni statističkom analizom su nesigurni jer se zasnivaju na promatranju samo manjeg dijela populacije, tj. na nepotpunim

podacima. Skup mjerenja na tom dijelu populacije proveden tijekom istraživanja nazivamo uzorak. Naš uzorak čini dio vozača iz već navedene populacije odabranih na slučajan način.

Cilj statističke analize je na osnovi podataka iz uzorka izvesti određene zaključke o populaciji te ocijeniti nesigurnosti koje su obuhvaćene tim zaključivanjem.

Za grafički prikaz podataka, kao i računanje konkretnih vrijednosti pri provođenju statističkih testova koristili smo se programom R [3].

2Opisna statistika

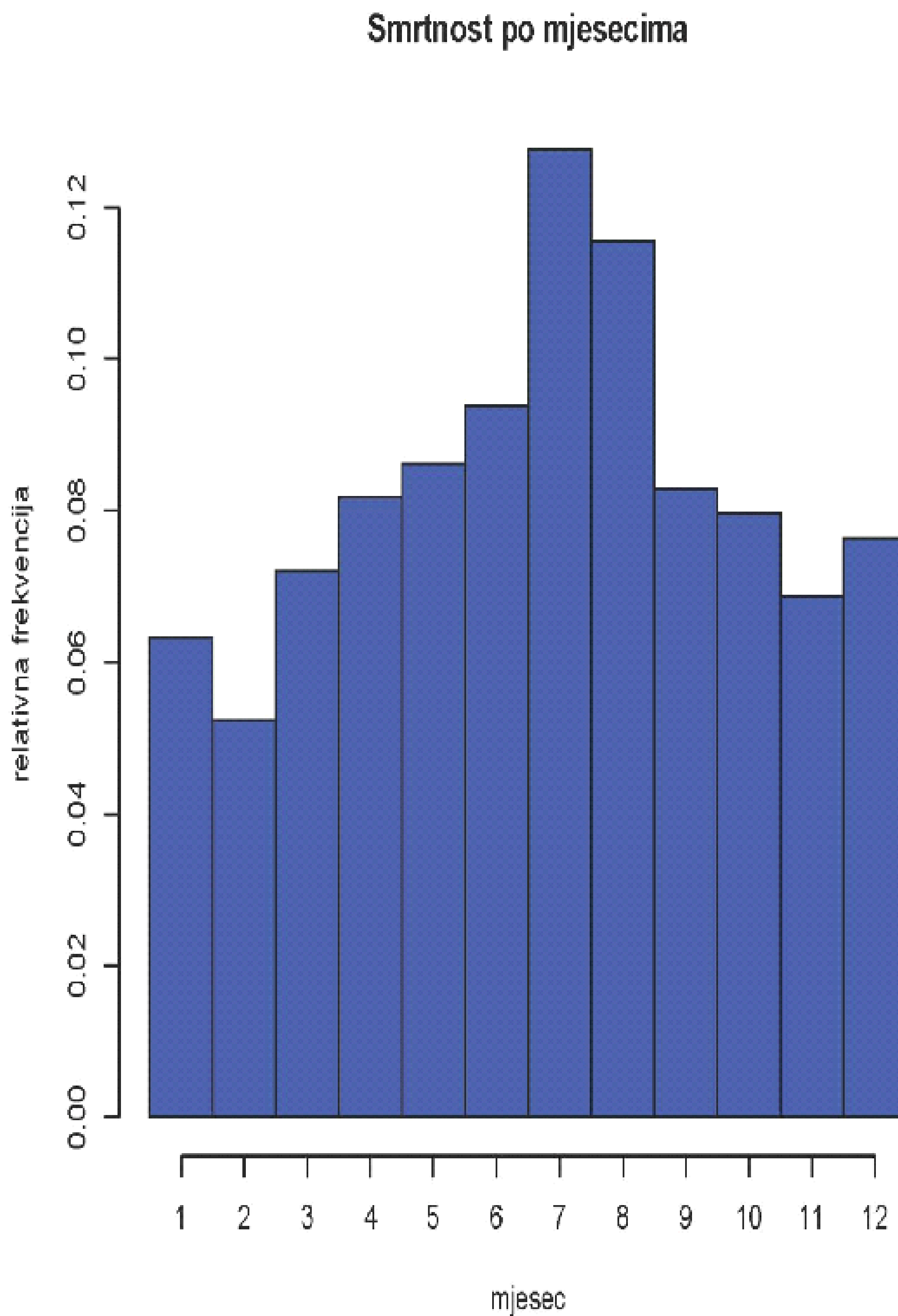
Opisna statistika je grana statistike koja se bavi predočavanjem i opisivanjem glavnih karakteristika prikupljenih podataka.

Za početak, korisno je podatke prikazati grafički, za što smo se koristili histogramima i strukturnim dijagramima.

Općenito, histogram je definiran kao način prikazivanja podataka raspoređenih u određene kategorije ili grupe. Kategorije, u koje smo grupirali podatke, nalaze se na osi apscisa, a prikupljeni podaci koji pripadaju određenoj kategoriji nalaze se na osi ordinata.

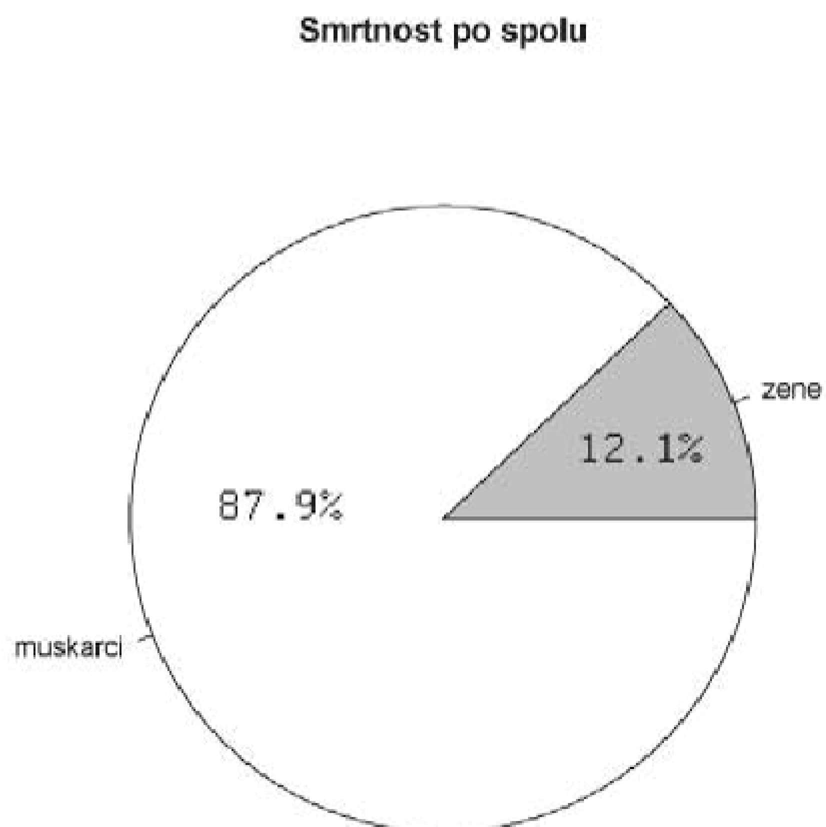
Kod strukturnog dijagrama svaka je kategorija ili grupa prikazana kružnim isječkom čija je površina proporcionalna udjelu te kategorije u uzorku.

Ovim izborom prikaza podataka dobiven je izvrstan pregled raspoređenosti broja nastradalih kroz mjesece u godini, te dobar uvid u spolnu i dobnu strukturu poginulih u promatranom razdoblju (slika 1).

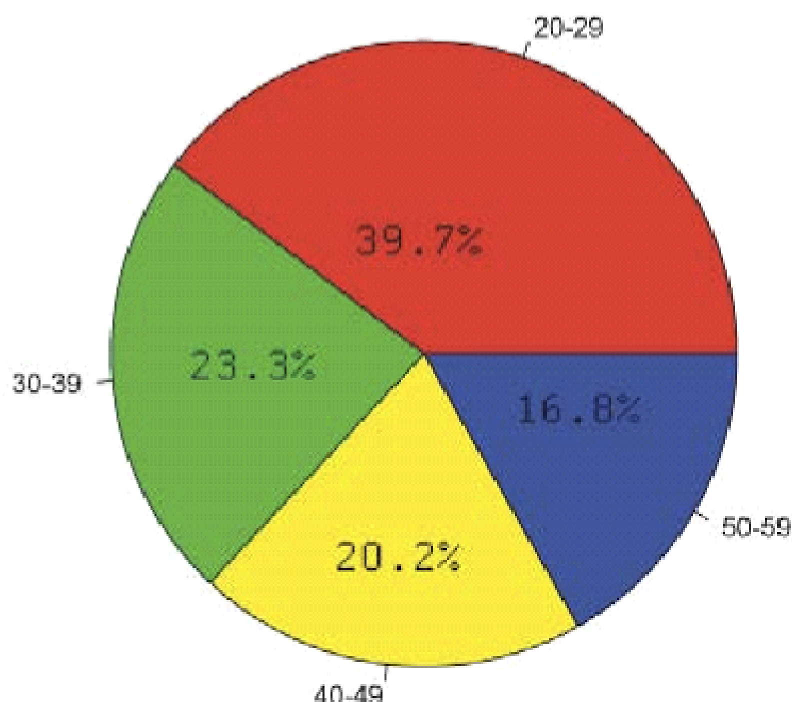


Slika 1: Histogram relativnih frekvencija broja poginulih tijekom 12 mjeseci

Iz histograma je očito da najviše ljudi pogine u srpnju, što je vjerojatno posljedica činjenice da tada najviše Hrvata kreće na godišnji odmor. Iako je uvriježeno mišljenje da su zimski mjeseci najopasniji za vozače zbog loših vremenskih uvjeta, iznenađujuće je da je najmanja smrtnost u siječnju i veljači.



Slika 2: Strukturni dijagram strukture poginulih



Slika 3: Strukturni dijagram strukture poginulih

Iz strukturnih dijagrama (slike 2 i 3) slijedi da najviše poginulih ima u dobnoj skupini od 20 do 29. Također možemo primijetiti da se broj poginulih smanjuje po dobnim skupinama, što govori da su stariji vozači oprezniji od onih u srednjim godinama, a oni u dobi od 20 do 29 najrizičnija su skupina.

Iako se za žene govori da su lošiji vozači od muškaraca, sa strukturnog dijagrama po spolu vidimo da pogine gotovo 7 puta više muškaraca nego žena.

3Testiranje statističkih hipoteza

Tijekom istraživanja mjeri se neko numeričko ili nenumeričko obilježje koje označavamo s XX. Rezultat mjerenja obilježja XX označavamo s xx. Slučajni uzorak tada možemo prikazati kao (X1,…,Xn)(X1,…,Xn), gdje je nn duljina uzorka, a s (x1,…,xn)(x1,…,xn) označiti jednu realizaciju tog uzorka.

Opažene frekvencije definiramo kao Nj=∑ni=11{Xi=aj}Nj=∑i=1n1{Xi=aj}, j=1,…,kj=1,…,k, pri čemu izraz 1{Xi=aj}1{Xi=aj} poprima vrijednost 1 ako je Xi=ajXi=aj, a inače poprima vrijednost 0, gdje je ajaj jedan od rezultata mjerenja obilježja XX u uzorku duljine nn.

Broj NjnNjn zove se relativna frekvencija.

Statistička hipoteza je bilo koja pretpostavka o distribuciji obilježja X, tj. pretpostavka da X ima sljedeću distribuciju:

$$\begin{pmatrix} a_1 & a_2 & \dots & a_k \\ p_1(\theta) & p_2(\theta) & \dots & p_k(\theta) \end{pmatrix},$$

pri čemu θ označava parametre o kojima ta distribucija može ovisiti, $a_1, a_2, \dots, a_k$ označavaju rezultate mjerenja, a $p_1(\theta), p_2(\theta), \dots, p_k(\theta)$ vjerojatnosti da će se ti rezultati postići.
S H_0 označavamo hipotezu koju želimo dokazati (to je tzv. nul-hipoteza), a s H_1 njoj alternativnu hipotezu.

Želimo na osnovi realizacije slučajnog uzorka za obilježje X donijeti odluku hoćemo li odbaciti hipotezu H_0 ili nećemo. Postupak donošenja odluke o odbacivanju ili neodbacivanju te statističke hipoteze zove se testiranje statističkih hipoteza.

Budući da su sve odluke bazirane na uzorcima koji nisu 100% pouzdani, niti zaključak statističkog testa nije 100% pouzdan. Test će biti potpun ako možemo procijeniti vjerojatnosti mogućih pogrešaka u zaključivanju. U većini slučajeva moguće je za zadanu razinu značajnosti testa α, $0<\alpha<1$, među testovima kojima vjerojatnost pogreške prve vrste ne prelazi broj α, naći test s najmanjom vjerojatnosti pogreške druge vrste. Pogrešku prve vrste radimo kad odbacujemo hipotezu H_0 i ona je istinita, a pogrešku druge vrste radimo kad zadržavamo hipotezu H_0 i ona je pogrešna (tj. hipoteza H_1 je istinita).

Kako na temelju dobivenih podataka i uz unaprijed određenu razinu značajnosti zaključiti odbacuje li se hipoteza H_0 i s kojom vjerojatnošću?

Prvo moramo izračunati vrijednost rezultata statističkog testa (test se odabire prema vrsti hipoteza), a zatim tu vrijednost usporediti s graničnom vrijednošću. Granična vrijednost je vrijednost testa za koju se hipoteza H_0 odbacuje, a ovisi o vrijednostima iz poznate distribucije vjerojatnosti specifične za odabrani test. Područje vrijednosti za koje se H_0 ne odbacuje nazivamo kritičnim područjem testa.

Jedan od najčešće korištenih testova u statistici je Pearsonov χ^2-test koji ćemo ovdje navesti, kako bi nam bio matematička podloga za daljnja istraživanja.

Definirajmo prvo očekivane frekvencije kao $n_j(\theta)=np_j(\theta), j=1,\dots,k$.

Neka je $D(\theta)=\sum_{i=1}^{k}\frac{(N_j-n_j(\theta))}{n_j(\theta)}^2$. Mi ćemo promatrati jednostavniji slučaj kada je hipotezom H_0 zadan parametar θ_0, čime je definirana testna statistika $H\equiv D(\theta_0)$.

Također definiramo broj stupnjeva slobode s $df=k-1$, a ako X ima χ^2-razdiobu, umjesto X pišemo $\chi^2(df)$. χ^2-razdioba je jedna od najčešćih razdioba u statistici i vrijednosti koje ona poprima zadane su tablično u tzv. tablici kvantila χ^2-razdiobe.

Sada smo spremni izreći već spomenuti Pearsonov teorem o χ^2-testu:

Ako je H_0 točna hipoteza, onda $H \xrightarrow{D} \chi^2(k-1)$, kada $n\to\infty$.

Za zadanu razinu značajnosti α, hipotezu H_0 odbacujemo ako je opažena vrijednost $h\geq\chi^2_\alpha(k-1)$, gdje vrijednost $\chi^2_\alpha(k-1)$ čitamo iz tablice kvantila χ^2-razdiobe.

S $\xrightarrow{D}$ označavamo konvergenciju po distribuciji, što jednostavnim rječnikom rečeno znači da se razdioba vrijednosti s lijeve strane približava razdiobi s desne strane kada $n \to \infty$. Često se koristi i oznaka $\sim$.

Pearsonov χ^2-test najčešće se upotrebljava ako je riječ o kvalitativnim podacima ili ako tim podacima distribucija značajno odstupa od normalne. Njegova primjena posebno se ističe u slučajevima kada želimo utvrditi odstupaju li dobivene frekvencije (iz slučajnog uzorka) od frekvencija koje bismo očekivali po hipotezi koju ispitujemo. Ovim testom također možemo ispitati povezanost dviju varijabli te vjerojatnost njihove povezanosti.

Općenito, χ^2-test najpouzdaniji je u sljedećim slučajevima:

(1)	Kada se ispituju odstupanja frekvencije uzorka od očekivane frekvencije uz zadanu hipotezu.
(2)	Kada se uspoređuju dva ili više nezavisnih uzoraka po nekom svojstvu, pri čemu su nam poznate frekvencije svakog od uzoraka.

3.1 Ovisnost smrtnosti u pojedninoj dobnoj skupini o spolu

Jedno od prvih pitanja koje nam se nametnulo pri proučavanju podataka jest jesu li spol vozača i njihova dob zavisna obilježja, tj. možemo li, s određenom sigurnošću, zaključiti da žene, odnosno muškarci imaju jednaku vjerojatnost pogibije u određenoj dobi. Možda naizgled ovo izgleda kao trivijalno, gotovo nevažno pitanje, no u statistici nas odgovori često mogu iznenaditi te ništa ne treba uzimati "zdravo za gotovo".

S obzirom na to da ovo ispitivanje spada u već navedene primjene χ^2-testa, odlučili smo se za njegovu varijantu χ^2-test nezavisnosti:

Promatramo dva različita obilježja X i Y. Neka je:

- n duljina uzorka,
- r broj različitih vrijednosti koje poprima obilježje X,
- c broj različitih vrijednosti koje poprima obilježje Y.

Neka je $((X_1,Y_1),\dots,(X_n,Y_n))$ slučajni uzorak iz dvodimenzionalnog statističkog obilježja (X,Y), pri čemu X može poprimiti vrijednosti $\{a_1,\dots,a_r\}$, a Y vrijednosti $\{b_1,\dots,b_c\}$.

χ^2-test nezavisnosti je statistički test kojim se testiraju hipoteze

H_0: X i Y su nezavisna obilježja H_1 : X i Y su zavisna obilježja

Po Pearsonovu teoremu, uz sitne promjene, možemo zaključiti da je testna statistika dana formulom

$$H=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(N_{ij}-n\hat{p}_i\hat{q}_j)^2}{n\hat{p}_i\hat{q}_j}\sim\chi^2(df),$$

gdje je

- N_{ij} opažena frekvencija od (a_i,b_j) u dvodimenzionalnom statističkom uzorku (X,Y),

··	$\hat{p}_i=\frac{N_i}{n}$, pri čemu je N_i opažena frekvencija od a_i u uzorku za X,
··	$\hat{q}_j=\frac{M_j}{n}$, pri čemu je M_j opažena frekvencija od b_j u uzorku za Y.

Područje $[\chi^2_\alpha(df),+\infty\rangle$, gdje je $df=rc-(r-1)-(c-1)-1$, nazivamo kritično područje. Ako je $h\in[\chi^2_\alpha(df),+\infty\rangle$, tada odbacujemo hipotezu H_0, a ako je h izvan tog intervala, onda je ne odbacujemo. Broj $\chi^2_\alpha(df)$ čitamo iz tablice kvantila χ^2-razdiobe.
U našem slučaju obilježje X (= spol) poprima vrijednosti muškarac, žena, a obilježje Y (= dobna skupina) poprima vrijednosti dobnih skupina, tj. 20 - 29, 30 - 39, 40 - 49, 50 - 59.

Podaci su prikazani sljedećom tablicom:

20 - 29	30 - 39	40 - 49	50 - 59	$\sum$	
Muškarac	327	186	161	131	805
Žena	37	28	24	23	112
$\sum$	364	214	185	154	917

χ^2-testom nezavisnosti koristimo se za testiranje sljedećih hipoteza:

H_0: Spol i dobna skupina su nezavisna obilježja H_1: Spol i dobna skupina nisu nezavisna obilježja

Test provodimo uz razinu značajnosti $\alpha=5\%$.
Račun provodimo u programu R [3]:

```
> x<-matrix(c(327,186,161,131,37,28,24,23),nrow=2,byrow=T)
> x
  [     [,1] [,2] [,3] [,4]]
  [[1,]  327  186  161  131]
  [[2,]  37   28   244   23]
> chisq.test(x)
  Pearson's Chi-squared test
  data:  x
  X-squared = 2.7395, df = 3, p-value = 0.4336
```

Odavde dobivamo da je $h=2.7395$ i $df=3$.
Promatramo u kojem intervalu se nalazi h. Budući da je $h<\chi^2_{0.05}(3)=7.8147$, tj. h nije unutar kritičnog područja, ne odbacujemo hipotezu H_0 i možemo zaključiti da su obilježlja X i Y nezavisna. Dakle, smrtnost u dobnim skupinama ne ovisi o spolu pa muškarci/žene imaju jednaku vjerojatnost da poginu u bilo kojoj starosnoj dobi.

3.2Ovisnost smrtnosti u pojedninoj dobnoj skupini o danima u tjednu

Jeste li se ikada zapitali pogine li više mladih vikendom ili u tjednu? Upravo nas je to potaknulo da provjerimo tvrdnju, često isticanu u medijima ,da najviše mladih nastrada u prometnim nesrećama tijekom vikenda.

Ponovo se koristimo χ^2-testom nezavisnosti, pri čemu obilježje X poprima vrijednosti dana u tjednu (ponedjeljak, utorak, srijeda, četvrtak, petak, subota i nedjelja), a obilježje Y poprima vrijednosti dobnih skupina, tj. 20 - 29, 30 - 39, 40 - 49, 50 - 59.

Podaci su prikazani sljedećom tablicom:

20 - 29	30 - 39	40 - 49	50 - 59	$\sum$	
Ponedjeljak	33	25	19	24	101
Utorak	27	25	35	18	105
Srijeda	37	25	22	18	102
Četvrtak	34	19	20	19	92
Petak	51	36	25	26	138
Subota	92	40	36	26	194
Nedjelja	90	44	28	23	185
$\sum$	364	214	185	154	917

Koristimo se χ^2-testom nezavisnosti (vidi 3.1) za testiranje sljedećih hipoteza:
H_0: Dan u tjednu i dobna skupina su nezavisna obilježja H_1: Dan u tjednu i dobna skupina nisu nezavisna obilježja

Test provodimo uz razinu značajnosti $\alpha=5\%$.
Računanjem u R-u [3], kod je vrlo sličan onome iz točke 3.1, dobiveni su sljedeći rezultati: $h=34.527$, $df=18$.

Budući da je $h>\chi^2_{0.05}(18)=28.8693$, odbacujemo hipotezu H_0 (jer se h nalazi u kritičnom području) i možemo zaključiti da obilježja X i Y nisu nezavisna. Dakle, smrtnost u dobnim skupina ovisi o danu u tjednu.

Budući da X i Y nisu nezavisna obilježja, sljedeće što nas zanima jest koliko jedno obilježnje ovisi o drugom. Konkretno, u našem slučaju, koliko su dobne skupine i dani u tjednu međusobno povezani. U statistici se ta povezanost mjeri stupnjem statističke zavisnosti koji je definiran formulom:

$$o=\frac{f^2}{\min\{r,c\}-1},$$

gdje je $f^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{N_{ij}}{N_i M_j}-1$ (za oznake vidi 3.1).
On je izračunat u R-u [3] i iznosi 1.27% pa je ta zavisnost veoma slaba, svakako slabija nego što bi to bilo za očekivati.

3.3 Ovisnost smrtnosti u pojedninoj dobnoj skupini o dobu dana

Sljedeće što ispitujemo jest distribucija smrtnosti po dobnim skupinama u određenom dijelu dana. Dijelove dana možemo promatrati kao nezavisne populacije pa se χ^2-test nameće kao logičan izbor. Ovu vrstu χ^2-testa u kojem se ispituje distribucija istog obilježja u više različitih uzoraka nazivamo χ^2-test homogenosti.

Pretpostavimo da nas zanima distribucija istog diskretnog statističkog obilježja X, koje poprima međusobno različite vrijednosti $\{a_1,\ldots,a_k\}$, u raznim populacijama.

Želimo na osnovi nezavisnih uzoraka uzetih iz tih populacija testirati nul-hipotezu da su razdiobe od X u tim populacijama jednake, tj. homogene.

Neka je m broj populacija. Iz svake populacije nezavisno odaberemo slučajni uzorak koji predstavlja obilježje X u i-toj populaciji i označimo ga s X_i, $i=1,\ldots,m$.

χ^2-test homogenosti je statistički test kojim se testiraju hipoteze

H_0: $X_1,\ldots,X_m$ su jednako distribuirani
H_1: postoje i i j takvi da se distribucija od X_i razlikuje od distribucije od X_j.

Po Pearsonovom teoremu slijedi da je testirana statistika dana formulom

$$H=\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(N_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}}\sim\chi^2(df),$$

gdje je

- N_{ij} opažena frekvencija od a_i u uzorku X_i,
- $\hat{n}_{ij}=\frac{n_i M_j}{n}$, $n_i=\sum_{j=1}^{k}N_{ij}$, $M_j=\sum_{i=1}^{m}N_{ij}$, $n=\sum_{j=1}^{k}M_j$.

Područje $[\chi^2_\alpha(df),+\infty\rangle$, gdje je $df=(m-1)(k-1)$, je kritično područje. Ako je $h\in[\chi^2_\alpha(df),+\infty\rangle$, tada odbacujemo hipotezu H_0, a ako je izvan tog intervala, onda je ne odbacujemo. Broj $\chi^2_\alpha(df)$ čitamo iz tablice kvantila χ^2-razdiobe.

Podaci su dani sljedećom tablicom:

20 - 29	30 - 39	40 - 49	50 - 59	$\sum$	
⟨0-6]	133	52	18	17	220
⟨6-12]	44	28	46	44	162
⟨12-18]	68	61	59	55	243
⟨18-24]	119	74	62	37	292
$\sum$	364	215	185	153	917

Koristimo se χ^2-testom homogenosti da bismo testirali hipotezu:
H_0: smrtnost u svakom promatranom dijelu dana jednako je distribuirana

Naš test ćemo provesti uz razinu značajnosti $\alpha=5\%$.
Računanjem u R-u [3] dobiveni su sljedeći rezultati: $h=94.7825$, $df=9$.

Iz danih podataka vidimo da je $h>\chi^2_{0.05}(9)=16.91898$, tj. h je unutar kritičnog područja, odbacujemo hipotezu H_0 i zaključujemo da smrtnost po dobima dana nije jednako distribuirana.

3.4 Utjecaj obrazovanja na smrtnost u svim dobnim skupinama

Proučavanjem podataka, nametnulo nam se pitanje ima li stupanj obrazovanja utjecaj na smrtnost u svim dobnim skupinama, pa smo odlučili provjeriti tu pretpostavku na vozačima sa završenom samo srednjom školom, tj. željeli smo odrediti postotak p takvih vozača u ukupnoj populaciji poginulih.

Za razliku od prijašnjih testova, sada ne uspoređujemo nekoliko populacija, već provjeravamo svoju pretpostavku unutar jedne populacije, pri čemu podatke tumačimo u odnosu na neko zadano obilježje (kod nas: završena samo srednja škola). To, naravno, znači da nam je potrebna drugačija testna statistika koja će nekako "odrediti" očekivani broj poginulih vozača sa završenom samo srednjom školom.

Kao i prije, ideja je pronaći takvu testnu statistiku koja će naše podatke svesti na neku nama poznatu distribuciju iz koje ćemo poslije lako pročitati s kojom vjerojatnošću smo postavili točnu hipotezu. Ovdje smo se poslužili poznavanjem Centralnog graničnog teorema, iz kojeg se odmah nametnula tražena statistika.

Navodimo Centralni granični teorem (CGT), kojim ćemo se poslije nekoliko puta koristiti:

Neka je $(X_n : n \in \mathbb{N})$ niz nezavisnih, jednako distribuiranih slučajnih varijabli s očekivanjem μ i varijancom σ^2, $0 < \sigma^2 < +\infty$, te neka je $T_n = \sum_{k=1}^{n} X_k$. Tada vrijedi $\frac{T_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{D} N(0,1)$ kad $n \to \infty$.

Iako smo CGT naveli u općenitom slučaju, nas zanima nešto jednostavnija situacija. Slučajni uzorak poginulih vozača možemo promatrati kao niz nezavisnih jednako distribuiranih Bernoullijevih slučajnih varijabli koje poprimaju vrijednost 0 ili 1 u ovisnosti o nekom zadanom svojstvu, i to s vjerojatnošću p, odnosno $1-p$.

Konkretno, mi ćemo svakog poginulog vozača koji ima završenu najviše srednju školu reprezentirati jedinicom u uzorku, dok će ostali biti reprezentirani nulom. Ovako promatran niz varijabli ima nešto jednostavnije formule varijance $(\sigma^2 = p(1-p))$ i očekvivanja $(\mu = p)$, pa je i testna statistika nešto jednostavnija nego u općenitom Centralnom graničnom teoremu. Također, sada je jasno da zapravo tražimo vjerojatnost p, tj. vjerojatnost da je poginuli vozač u uzorku imao završenu samo srednju školu.

Test ovoga oblika, u kojem računamo očekivanje za populaciju reprezentiranu Bernoullijevim varijablama, nazivamo Z-test i definiramo testnu statistiku (s opravdanjem u CGT-u i jer je $n\bar{X}_n = T_n$) formulom:

$Z = \frac{\bar{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$.

Ovo je najjači test za računanje očekivanja uz razinu značajnosti α , gdje je

- n duljina uzorka
- $\bar{X}_n$ relativna frekvencija vozača sa završenom samo srednjom školom u uzorku.

Promatrajući svoje podatke, uočili smo da najveći broj poginulih vozača ima završenu samo srednju školu pa smo opisanim Z-testom odlučili provjeriti svoje očekivanje da takvih vozača ima otprilike 70%.

Ovdje je važno napomenuti da je statističko istraživanje često puno pretpostavki dobivenih tzv. "metodom pokušaja i pogrešaka", te često nije moguće iz prve pogoditi koja je hipoteza optimalna.

Dakle, testirat ćemo sljedeće hipoteze:

H_0: $p=0.70$ H_1: $p>0.70$.

Test ćemo provesti uz razinu značajnosti $\alpha=5\%$.
Za podatke dobivamo $\overline{X}_n=0.7388$.

Uvrštavanjem konkretnih vrijednosti iz uzorka duljine $n=781$ dobivamo sljedeće:

$$z=\frac{0.7388-0.7}{\sqrt{\frac{0.7\cdot 0.3}{781}}}=2.3662>z_{0.05}=1.64,$$

gdje broj $z_{0.05}$ čitamo iz tablice standardne normalne distribucije ($z_{0.05}=\Phi(1-0.05)$).
Promatramo u kojem intervalu se nalazi z. Ako je $z\in[z_{0.05},+\infty\rangle$, tada odbacujemo hipotezu H_0, u protivnom je ne odbacujemo.

Dobiveni rezultat je iz intervala $[z_{0.05},+\infty\rangle$ pa odbacujemo hipotezu H_0 u korist hipoteze H_1 i možemo zaključiti da više od 70% poginulih ima završenu samo srednju školu.

3.5 Očekivana dob vozača u trenutku nesreće

Pitanje koje se prirodno nameće je očekivana dob u trenutku nesreće. Točnije, zanima nas možemo li pronaći neki interval godina vozača u kojem je vjerojatnost nesreće najveća. U statistici takav interval nazivamo aproksimativni pouzdani interval.

Prema CGT teoremu znamo da je $Z=\frac{\overline{X}_n-\mu}{\frac{\sigma}{\sqrt{n}}}\sim N(0,1)$ za velike n.

Po formuli za vjerojatnost vrijedi: $P(|Z|\leq z_{\frac{\alpha}{2}})=1-\alpha$,

što je ekvivaletno s $P(z_{\frac{\alpha}{2}}\leq\frac{\overline{X}_n-\mu}{\frac{S_n}{\sqrt{n}}}\leq z_{\frac{\alpha}{2}})=1-\alpha$,

što je ekvivaletno s
$P(\overline{X}_n-z_{\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}\leq\mu\leq\overline{X}_n+z_{\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}})=1-\alpha$.

Dakle, interval je dan formulom

$$\left[\overline{X}_n-z_{\frac{\alpha}{2}}\cdot\frac{S_n}{\sqrt{n}},\overline{X}_n+z_{\frac{\alpha}{2}}\cdot\frac{S_n}{\sqrt{n}}\right],$$

gdje je

- n duljina uzorka,
- x_i godine života i-te osobe u trenutku nesreće,
- $\overline{X}_n=\frac{\sum_{i=1}^{n}x_i}{n}$,
- $S_n^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{X}_n)^2$ procjenitelj za varijancu,

a broj $z_{\frac{\alpha}{2}}$ čitamo iz tablice standardne normalne distribucije.

Iz uzorka duljine $n=917$ dobivamo $\overline{X}_n=35.46$ i $S_n=11.72$. Dakle, aproksimativni 95% pouzdani interval za očekivanu dob u trenutku pogibije je $[34.7,36.22]$ pa zaključujemo da je očekivana dob između 34 i 37 godina.

3.6Utjecaj promjene Zakona o sigurnosti prometa na cestama na smrtnost u dobnoj skupini od 20 - 29

Iz strukturnog dijagrama o udjelu pojedinih dobnih skupini u ukupnom broju poginulih, vidjeli smo da je najugroženija skupina u dobi od 20 do 29 godina. S obzirom na to da se i Zakon o sigurnosti prometa na cestama u 2008. bazirao upravo na toj dobnoj skupini, odnosno mladim vozačima, želimo utvrditi je li on uistinu utjecao na smanjenje smrtnosti.

Promatramo podatke o poginulima u toj dobnoj skupini u razdoblju od godine dana nakon donošenja prvog zakona u 2004. godini (prvo razdoblje) te od godinu dana nakon donošenja novog zakona u 2008. godini (peto razdoblje).

Pretpostavljamo da novi zakon ima manji utjecaj na smrtnost u dobnoj skupini od 20 do 29 od starog pa želimo naći neki test kojim bismo mogli usporediti "uspješnost" ovih dvaju zakona. Za početak, tu "uspješnost" zakona definiramo kao udio poginulih vozača u dobi od 20 do 29 godina u cjelokupnom broju poginulih. Sada je još potrebno naći najbolji način da usporedimo omjere prvog i petog razdoblja.

Odabrali smo test omjera proporcija koji se koristi upravo u situacijama kada uspoređujemo "uspješnost" nekog obilježja u nezavisnim populacijama.

Test omjera proporcija provodi se na dvije nezavisne populacije s nekim obilježjem X.

Označimo s X_1 slučajnu varijablu koja predstavlja obilježje X u prvoj populaciji, a s X_2 slučajnu varijablu koja predstavlja X u drugoj populaciji.

Neka su p_1 i p_2 njihove vjerojatnosti uspjeha u svakoj od populacija.

U osnovnoj nul-hipotezi pretpostavljamo da su vjerojatnosti uspjeha jednake, a druga hipoteza je njena alternativa koja ovisi o zadatku.

Test omjera proporcija definiran je formulom:

$$Z=\frac{\hat{p}_2-\hat{p}_1}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}},$$

gdje su n_1 i n_2 dovoljno velike populacije (zbog CGT-a), $\hat{p}_1$ procjenitelj za p_1 (tj. $\hat{p}_1=p_1$), $\hat{p}_2$ procjenitelj za p_2 (tj. $\hat{p}_2=p_2$) i $\hat{p}=\frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2}$ procjenitelj zajedničke vjerojatnosti.
U našem slučaju promatrano obilježje je smrtnost, a populacije su poginuli u prvom i petom razdoblju. Označimo vjerojatnosti s p_1 = omjer poginulih u dobi od 20 do 29 u prvom razdoblju, p_5 = omjer poginulih u dobi od 20 do 29 u petom razdoblju.

Testiramo sljedeće hipoteze uz razinu značajnosti $\alpha=5\%$:

H_0: $p_1=p_5$ H_1: $p_1<p_5$

Koristeći se navedenim formulama, za svoje podatke dobivamo ove rezultate:

·· $n_1=157$
·· $n_5=195$
·· $\hat{p}_1=\frac{56}{157}=0.3567$

- $\hat{p}_5=\frac{74}{195}=0.3795$
- $\hat{p}=0.3693$
- $Z=0.4406<z_{0.05}=1.64$,

gdje broj $z_{0.05}$ čitamo iz tablice standardne normalne distribucije.

Promatramo u kojem intervalu se nalazi z. Ako je $z\in[z_{0.05},+\infty\rangle$, tada odbacujemo hipotezu H_0, u protivnom je ne odbacujemo. Budući da z nije iz tog intervala, ne možemo odbaciti hipotezu H_0, odnosno novi i stari zakon imaju jednak utjecaj na smrtnost u dobnoj skupini od 20 do 29.

4 Zaključak

Istaknimo na kraju najzanimljivije rezultate rada:

- unatoč uvriježenoj pretpostavci, žene nisu lošiji vozači od muškaraca, štoviše, gotovo sedam puta manje žena pogine u prometnim nesrećama
- smrtnost mladih ovisi o danu u tjednu
- više od 70% poginulih ima završenu samo srednju školu
- očekivana dob u trenutku pogibije je između 34 i 37 godina
- promjena Zakona o sigurnosti prometa na cestama nije utjecala na smanjenje smrtnosti mladih.

5 Literatura

Bibliografija

[1] M. Huzak - Predavanja iz statisike

[2] N. Sarapa - Teorija vjerojatnosti, Školska knjiga, Zagreb, 1992.

[3] R

[4] Wikipedijini članci o histogramu i Z-testu

[5] MUP